TU Dortmund

Introductory Case Studies

Project 1: Descriptive Analysis of Demographic data

Lecturers:
Prof. Dr. Crystal Wiedner
Prof. Dr. Rouven Michel
Prof. Dr. Marlies Hafer

Author: Siddhartha Karki (Matr. No : 238092)

Group number: 4

Group members:
Sagar Basnet

November 3, 2024

Contents

1 Introduction

Demographic data is crucial for understanding population characteristics and guiding policy on global challenges related to health, economic stability, and social planning. Specifically, life expectancy and under-5 mortality rates are key indicators of health outcomes and living standards. The International Database (IDB), managed by the U.S. Census Bureau, provides a comprehensive and regularly updated resource for global demographic data from 227 countries. This database encompasses demographic variables from 1950 to projected values for 2100, offering an extensive view of global population dynamics. For this project, a subset of IDB data from the Statistical Department of TU Dortmund includes life expectancy and under-5 mortality rates for 2004 and 2024, categorized by sex across five regions and 21 subregions, enabling focused analysis of trends in these indicators.(Bureau, 2022).

Descriptive analysis summarizes data to reveal key characteristics and identify patterns and trends. The goal of this project is to conduct such an analysis on the dataset using measures of central tendency, variability, and correlation to explore relationships and distributions. Frequency distributions are examined through histograms and mean calculations, while the Pearson correlation coefficient assesses relationships between variables and graphically displayed by using correlation heat map. Interquartile ranges, shown with box plots, highlight demographic homogeneity and heterogeneity with reference to Europe region. To perform frequency distribution and bivariate correlation analysis while assessing variable homogeneity we used 2024 data from dataset. Finally, the changes between 2004 and 2024 data are illustrated with histograms.

In section 2 an overview on methods of data collection, size of data, description of variables and the quality of data is highlighted. In section 3 the statistical methods and the tools used for the implementation of the methods are explained. Here, the descriptive statistical methods that have been used in this project the methods like means, median, interquartile range, and correlation coefficients are clearly explained. In section 4 interpretation of findings after the implementation of statistical methods with the help of graphical plots such as histograms, scatter plots, and box plots is carried out. Finally, the conclusion of finding , summary of the report and possible further investigation is discussed in section 5.

2 Problem Statement

2.1 Data Set and Data Quality

The dataset for this project is sourced from the International Database (IDB) published by the U.S. Census Bureau. It includes ten variables from 227 countries for 2004 and 2024, with life expectancy and under-5 mortality rates stratified by sex, and countries organized into five regions and 21 subregions, as detailed in Table 1. During inspection, five instances of special characters were found in the "Life Expectancy" and "Under-5 Mortality Rate" variables, leading to their removal to maintain data quality. These unidentified values constitute a small portion of the dataset and are unlikely to significantly affect statistical findings; overall, the data is considered reliable due to its official sources.

Table 1: Characteristics of Variables

Variable Name	**Data Type**	**Description**
Region	Categorical	Includes five different continents.
Subregion	Categorical	Represents part of a continent based on geographic location.
Life expectancy at birth of both sexes	Numeric	Refers to the predicted number of years a newborn of any sex is expected to live.
Life expectancy at birth of males	Numeric	Refers to the predicted number of years a newborn male is expected to live.
Life expectancy at birth of females	Numeric	Refers to the predicted number of years a newborn female is expected to live.
Under age-5 mortality rate of both sexes	Numeric	The average number of infants dying before reaching 1 year of age
Under age-5 mortality rate of males	Numeric	Refers to the number of deaths of male children under five years of age per 1,000 live male births in a given year.
Under age-5 mortality rate of females	Numeric	Refers to the number of deaths of female children under five years of age per 1,000 live female births in a given year.
Total Fertility Rate	Numeric	The total number of children that would be born to each woman.

3 Project Objective

The primary objective of this project is to conduct a descriptive analysis of the provided demographic data. The analysis begins by assessing the frequency distribution of the variables using statistical techniques such as mean, quartiles, minimum, and maximum values, illustrated through histograms and box plots. Next, bivariate correlations among the variables are examined using Pearson's Correlation Coefficient to evaluate individual relationships. The study also investigates homogeneity within and between subregions by calculating measures of central tendency. Finally, a comparative analysis of data from 2004 and 2024 identifies changes in variable values over time, with findings presented in histogram.

4 Statistical Methods

4.1 Central Tendency

In statistics, central tendency is a value used to represent an entire dataset, providing an overview of where most values lies. There are three main measures of central tendency: the mean, the median, and the mode. (Bhandari, 2023). It provides insight into the overall distribution of values in a dataset. Central tendency works alongside measures of variability (which describe how spread out the values are) as a core element of descriptive statistics.

For the given case study, mean and median is used for the finding the frequency distribution of variables.

4.1.1 Mean

Mean is a key measure of central tendency. It is calculated by adding up all the values in a dataset and then dividing the sum by the number of values. The mean is particularly useful in datasets where values are evenly distributed, as it provides a reliable snapshot of the data's overall trend. The mean (arithmetic mean) calculated for sample data is denoted by $\overline{x}$ (Taylor, 2022).

Let $x_1, x_2, \ldots, x_n$ be 'n' number of data points then, mean is given by:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $x_1, x_2, \ldots, x_n$ are sample data from 'n' observed data.

4.1.2 Median

The median of a dataset is the middle value when the data is ordered from lowest to highest. In larger datasets, formulas are often used to locate the position of this middle value more efficiently. The method for finding the median differs depending on whether the dataset has an odd or even number of values. (geeksforgeeks, 2022).

For an odd-numbered dataset, the position of the median is:

$$\text{Position} = \frac{n+1}{2}$$

For an even-numbered dataset, the median is found by taking the average of the two middle values, at the following positions:

$$\text{Positions} = \frac{n}{2} \quad \text{and} \quad \left(\frac{n}{2}+1\right)$$

To calculate the median from the two middle values:

$$\text{Median} = \frac{\text{Value at } \frac{n}{2} + \text{Value at } \left(\frac{n}{2}+1\right)}{2}$$

4.2 Variance

Variance is a statistical measure that quantifies the degree of spread or dispersion in a set of data. Variance helps us understand the extent to which data points deviate from the central tendency. Here, variance is used for finding the homogeneity between the individual variables within the subregions (WallstreetMojo, 2021).It is denoted by σ^2 and is mathematically given by :

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{N}$$

where 'N' is total number of values in data set, $\overline{x}$ is mean and x_i is the values in data set.

4.3 Quartiles

Quartiles are values that partition a dataset into four equal segments. They provide a quick way to assess the spread and central tendency of the data, which are essential initial steps in data analysis. (Liberto, 2022).

- The first quartile (Q1) is defined as the 25th percentile, where the lowest 25% of the data falls below this point. It is also known as the lower quartile.
- The second quartile (Q2) is the median of a dataset; thus, 50% of the data lies below this point.
- The third quartile (Q3) is the 75th percentile, where the lowest 75% of the data falls below this point. It is known as the upper quartile, as 75% of the data lies below this point.

4.4 Inter-Quartile Range (IQR)

The Interquartile Range (IQR) is a valuable measure in descriptive analysis. It provides insights into the spread of data within the middle half of a distribution. It is calculated by finding the difference between the third quartile (Q3) and the first quartile (Q1), representing the range of the central 50% of the data. (Frost, 2022)

It is denoted by IQR and is given by:

$$IQR = Q_3 - Q_1$$

where, Q_1 (first quartile) indicates the how the data is distributed in first 25% of distribution and Q_3 (third quartile) indicates how the data is distributed in 75% of distribution.

4.5 Correlation Coefficient

A correlation coefficient is a numerical value ranging from -1 to 1 that indicates the strength and direction of a relationship between variables. It essentially shows how closely related the measurements of two or more variables are within a dataset. This coefficient is represented by a numerical value that falls within the range of +1 to -1. A value nearing +1 implies a positive relationship between the variables, which means that as one variable increases, the other tends to increase as well. Conversely, a value

approaching -1 suggests a strong negative relationship, indicating that as one variable increases, the other tends to decrease. On the other hand, a correlation coefficient close to 0 signifies a lack of a substantial relationship between the variables. (Bhandari, 2022).

4.5.1 Pearson's Correlation Coefficient

Pearson's Correlation Coefficient is a widely utilized method for assessing the correlation between two variables. This coefficient provides insight into whether there exists a linear relationship between the variables within a dataset. Represented by a value ranging from +1 to -1, Pearson's Correlation Coefficient quantifies both the strength and direction of the linear association between variables (Bhandari, 2022). It is denoted by "r" and is given by:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

Where:

$$\begin{aligned} r &: \text{Pearson correlation coefficient} \\ X_i &: \text{Individual data points of the first variable } X \\ \bar{X} &: \text{Mean of the variable } X \\ Y_i &: \text{Individual data points of the second variable } Y \\ \bar{Y} &: \text{Mean of the variable } Y \end{aligned}$$

4.6 Graphical Method

In this section, we discuss about the graphs and plot used in the case study to explain the solution of found problems.

4.6.1 Histogram

A histogram is a graphical tool for visualizing the probabilistic distribution of numerical or image data. Specifically designed for understanding the distribution of a continuous variable, a histogram is created by defining bins or classes for numerical data. Once

these bins are established, the histogram counts the occurrences of values within each bin and illustrates this information in a graph. The bars on the graph represent the frequency of values within each bin, offering a visual depiction of the distribution across the dataset (Yi, 2019).

4.6.2 Box plot

The box plot, also known as a box-and-whisker plot, is a graphical method for visually presenting the distribution of data based on key statistical indicators: the minimum value, the first quartile (Q1), the median, the third quartile (Q3), and the maximum value. This visual tool not only communicates the symmetry or skewness of the data but also highlights how the data is grouped. The rectangular "box" spans from Q1 to Q3, with the median represented by a line inside, and "whiskers" extending to the minimum and maximum values (Lane, 2001).

4.6.3 Scatter plot

A scatter plot serves as a graphical representation depicting the relationship between two variables. With variables assigned to the x-axis and y-axis, each data point is plotted as (x, y). This visual tool is used in illustrating the correlation between the two variables. A positive correlation is evident when the values of x increases as increase in the values of y. Conversely, a negative correlation is observed when the values of x increase while the values of y decrease. If no clear relationship is seen, it indicates zero correlation (Wilkes, 1994).

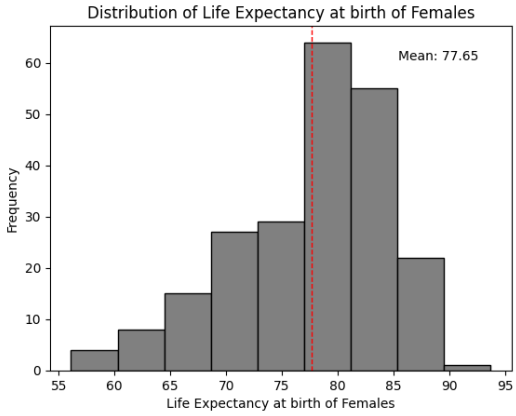
5 Statistical Analysis

In this section, the results obtained through statistical methods are visualized and interpreted. The analysis was conducted using the Python programming language along with several of its built-in libraries. The software requirements for this project include Python (version 3.10.4), along with the libraries NumPy (version 1.20.3), Pandas (version 1.3.4), and Seaborn (version 0.13.0).

5.1 Frequency Distribution of variables

This case study includes histogram to show how the frequency distribution of given variables looks like.

(a) Life expectancy at birth of male

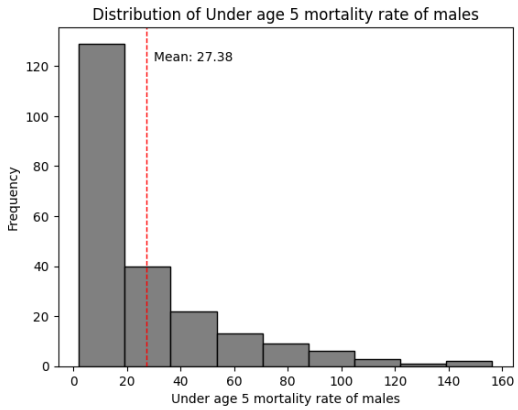


(b) Life expectancy at birth of female

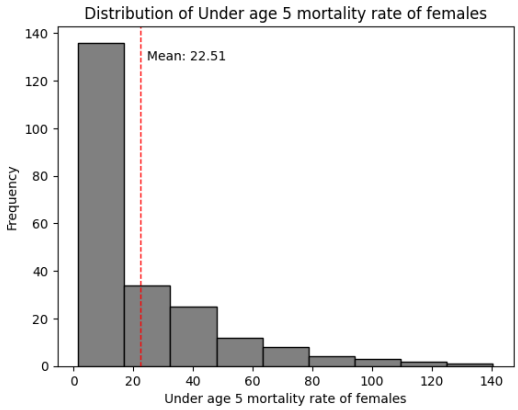
(c) Life expectancy at birth of both sexes

Figure 1: Frequency distribution of variables

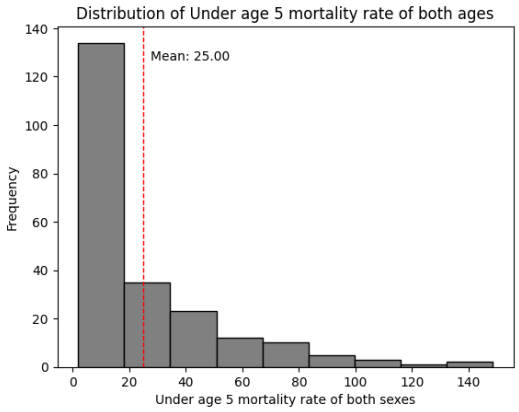
Figures 1(a), 1(b), and 1(c) depict the distribution of life expectancy at birth for males, females, and both sexes combined. In Figure 1(a), the average life expectancy for males is 72.66 years, with nearly 160 countries reporting life expectancies between 70 and 80 years; the highest life expectancy recorded is 85 years, while the lowest is 55 years. Figure 1(b) shows that the average life expectancy for females is 77.65 years, with a maximum of 95 years and a minimum of 57 years. Finally, Figure 1(c) presents the frequency distribution of life expectancy for both sexes, revealing an average of 75.10 years, with only two countries reaching a life expectancy of 90 years.



(a) Under age-5 mortality rate of males



(b) Under age-5 mortality rate of males



(c) Under age-5 mortality rate of both sexes

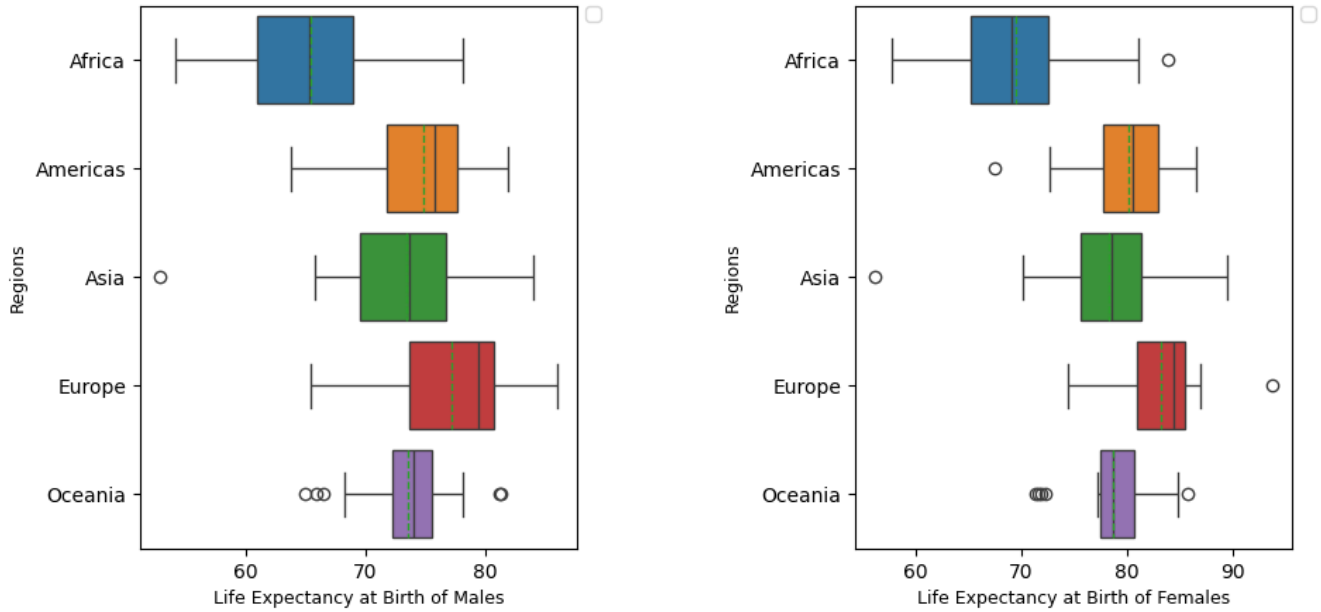
Figure 2: Frequency distribution of variables

Figure 2 illustrates the under-5 mortality rates by gender and combined view: Figure 2(a) shows the male under-5 mortality rate, with a global average of approximately 27.38 deaths per 1,000 live births. Around 125 countries report male mortality rates between 0 and 20 per 1,000, reflecting relatively low mortality across much of the world, though about three countries have exceptionally high rates reaching up to 160 deaths per 1,000. Figure 2(b) depicts the female under-5 mortality rate, with a global average of 22.51 deaths per 1,000, and roughly 155 countries in the 0–20 range, indicating generally low female mortality. However, about 10 countries exhibit significantly higher rates for females, ranging from 100 to 140 per 1,000. Figure 2(c) presents the combined under-5 mortality rate for both sexes, with a global average of approximately 25 deaths

per 1,000. Nearly 130 countries fall within the 0–20 range, yet about three countries experience notably higher rates, reaching up to 140 deaths per 1,000.

Differences between the sexes and regions

In this subsection, it is shown that how the variables of two different genders are different within and between regions.

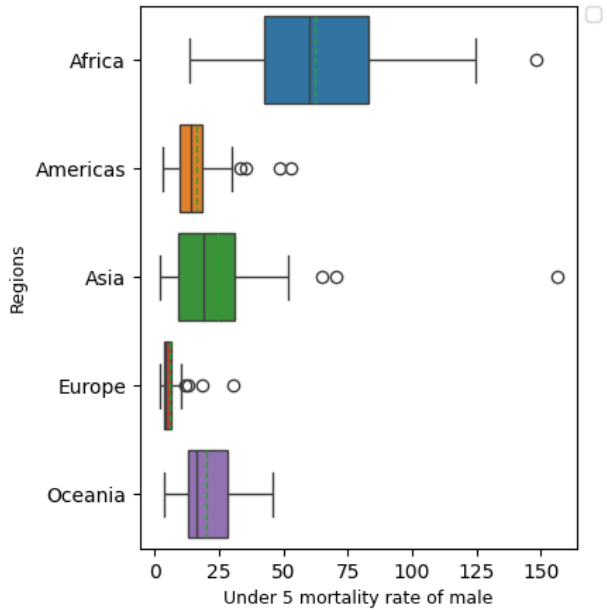


(a) Median age of males in different regions

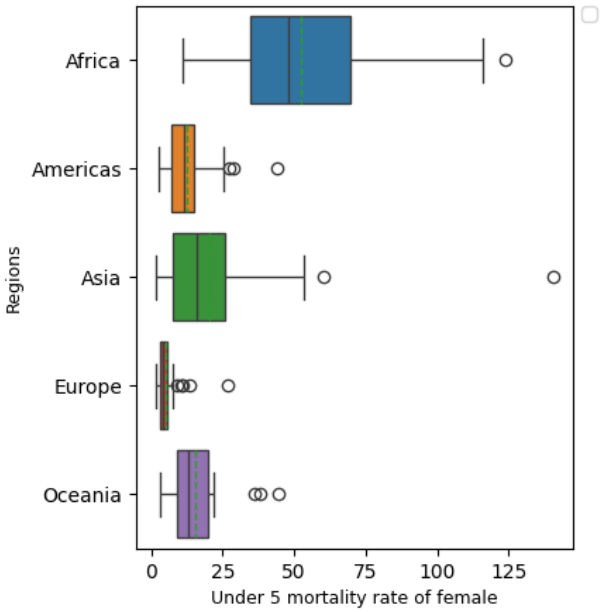
(b) Median age of females in different regions

Figure 3: Differences between median age of sexes and regions

Figure 3 illustrates life expectancy at birth by gender across different regions: Figure 3(a) shows that male life expectancy is highest in Europe, with an average of 78 years, significantly surpassing other regions. The Americas follow with a male life expectancy of 73 years, showing a moderate gap compared to Europe, while Asia ranks next at 71 years, indicating slightly lower longevity than the Americas. Africa has the lowest male life expectancy, averaging 65 years, reflecting a more considerable disparity compared to other regions. Figure 3(b) highlights female life expectancy, which is also highest in Europe at approximately 83 years, followed by the Americas at around 80 years and Asia at an average of 78 years. Africa has the lowest female life expectancy at birth, at about 70 years.



(a) Under-5 infant mortality rate males



(b) Under-5 infant mortality rate females

Figure 4: Differences between Under-5 infant mortality rate of sexes and regions

Figure 4 illustrates the under-5 mortality rate by gender across various regions: Figure 4(a) shows that the average under-5 mortality rate for males is highest in Africa, at 65 deaths per 1,000 live births, while Europe has the lowest rate, averaging around 5, highlighting a stark contrast in child mortality outcomes. Following Europe, the Americas have a slightly higher average of approximately 10, with Asia presenting a moderate average under-5 mortality rate of 25 and Oceania showing a similar average of about 23. Figure 4(b) illustrates the under-5 mortality rate for females, revealing that Africa again has the highest average at 55 deaths per 1,000 live births, contrasted by Europe with the lowest rate of around 3, indicating a significant difference between these regions. The Americas follow Europe with a low average of approximately 5, while Asia and Oceania have moderate averages of 24 and 22, respectively.

5.2 Variability of Variables within subregions

Variability in statistics pertains to how data points within a dataset differ from each other or from the mean. To analyze the homogeneity of variables across sub-regions, the region of Europe has been selected for focused examination. The dataset provides data for four distinct sub-regions within Europe.

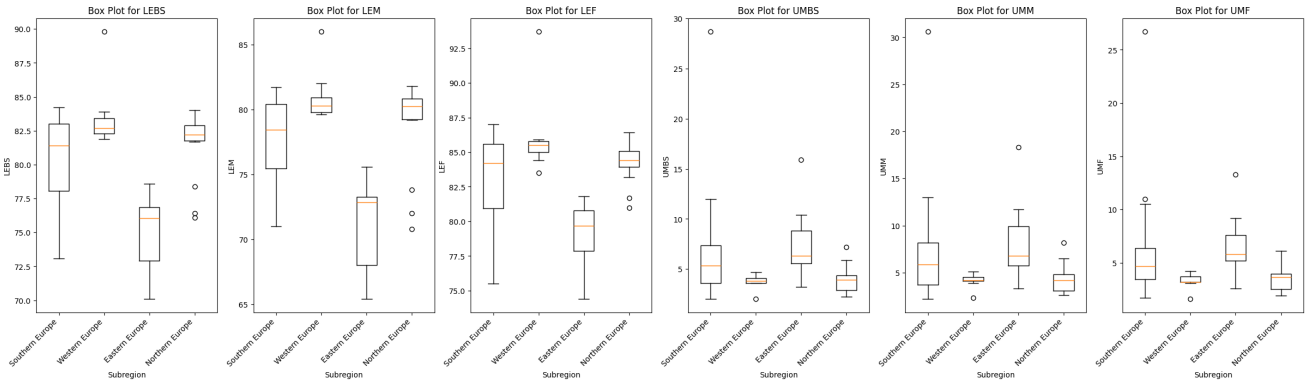


Figure 5: Life expectancy at birth of male

Figures 5(a), 5(b), and 5(c) demonstrate a consistent pattern of life expectancy across both sexes in Southern and Northern Europe, with median values closely aligned and interquartile ranges (IQRs) showing minimal variation: 4.92 for both sexes, 4.95 for males, and 4.65 for females in Southern Europe; and 1.12 for both sexes, 1.62 for males, and 1.125 for females in Northern Europe. Similarly, Figures 5(d), 5(e), and 5(f) indicate uniform under-5 mortality rates in Northern and Western Europe, with IQRs of 1.47 for both sexes, 1.69 for males, and 1.45 for females in Northern Europe, and even lower values in Western Europe. While life expectancy figures show significant regional disparities, reflecting various contributing factors, under-5 mortality rates remain stable across these sub-regions, indicating similar conditions affecting child health outcomes. This results in a consistent demographic characteristic for under-5 mortality despite the variability in life expectancy.

5.3 Bi-variate Correlation

Bivariate correlation analysis is a statistical method used to evaluate the relationship between two variables. Figure 6 illustrates strong positive associations among the life expectancy measures, with a correlation coefficient of 0.991986 between the life expectancy of both sexes (LEBS) and the life expectancy of males (LEM), and a coefficient of 0.992008 between LEBS and the life expectancy of females (LEF). Additionally, a significant correlation of 0.968196 exists between LEM and LEF, indicating that higher life expectancy for one gender is closely linked to higher life expectancy for the other. Conversely, the analysis reveals strong negative correlations between life expectancy and under-5 mortality rates, highlighting the inverse relationship between these variables. Specifically, LEBS exhibits a correlation of -0.889654 with the under-5 mortality rate

for both sexes (UMBS), suggesting that as life expectancy increases, the under-5 mortality rate tends to decrease. Similar patterns are evident with under-5 mortality rates for males (UMM) and females (UMF), which show strong negative correlations of -0.887954 and -0.888383, respectively.

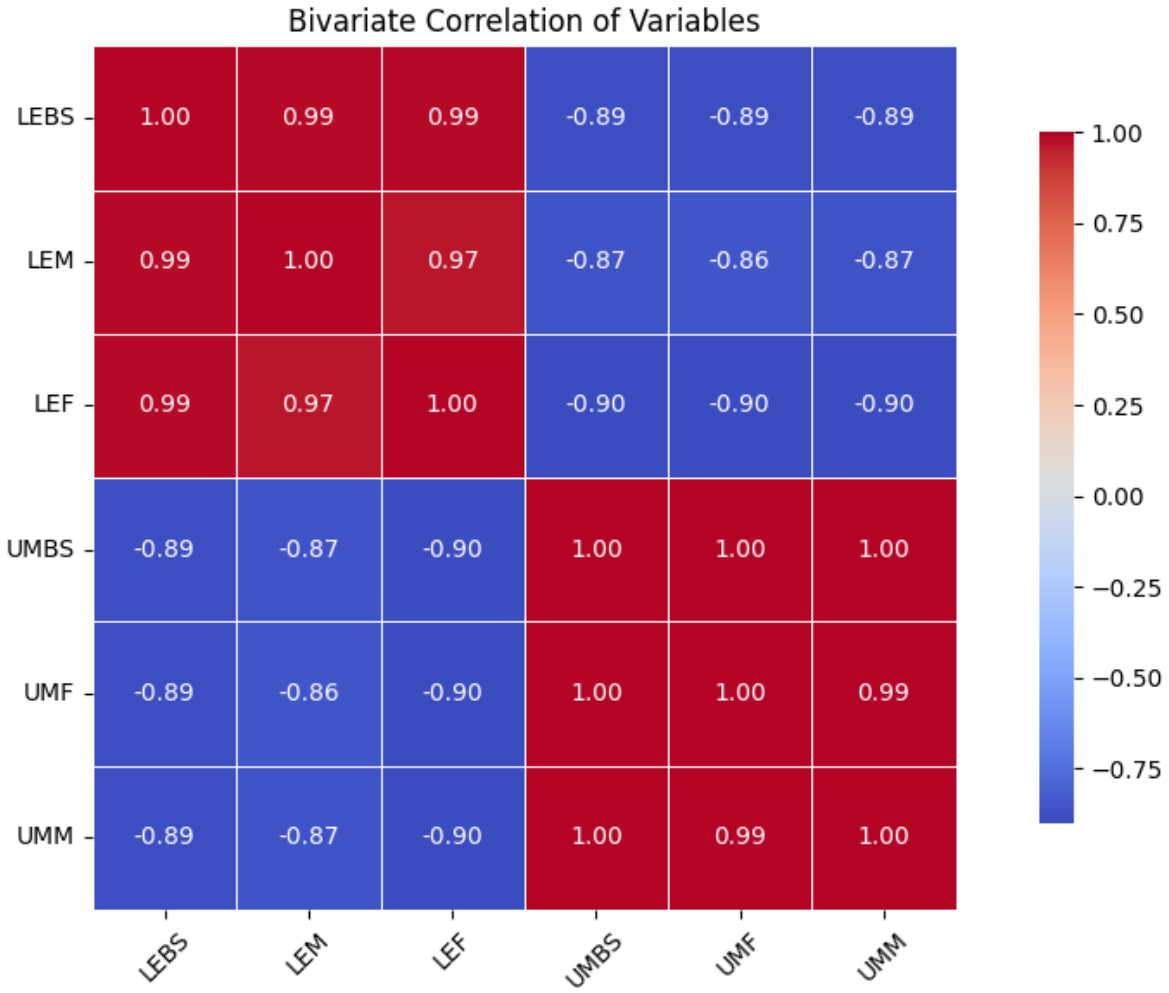


Figure 6: Bivariate correlation of variables

Furthermore, the under-5 mortality rates themselves demonstrate a high degree of interconnection, as evidenced by the strong positive correlations among them. The correlation between UMBS and UMF is particularly high at 0.997947, while UMBS and UMM exhibit an almost perfect correlation of 0.998570. These findings underscore the consistent trends in under-5 mortality rates across genders, suggesting shared factors influencing child health outcomes.

5.4 Changes in variable over 20 years

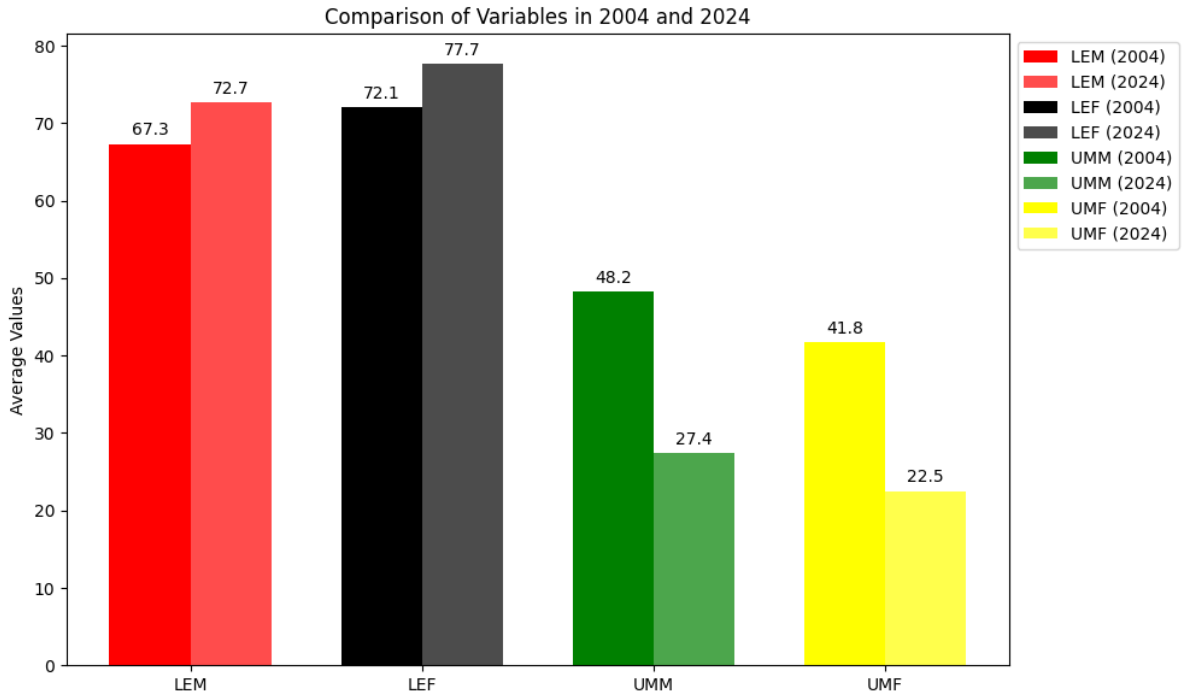


Figure 7: Comparison of Variables

The bar chart clearly illustrates that there has been a remarkable increase in life expectancy for both males and females, with males gaining 5.4 years, rising from 67.3 years in 2004 to 72.7 years in 2024. Females experienced an even more significant improvement, with life expectancy increasing by 5.6 years, from 72.1 years to 77.7 years. Additionally, the data reveals a dramatic reduction in under-5 mortality rates for both genders. Male under-5 mortality rates fell significantly by 20.8, from 27.4 to 6.6, while female under-5 mortality rates decreased by 19.3, dropping from 22.5 to 3.2.

6 Summary

This project analyzed demographic data from the International Database of the U.S. Census Bureau, focusing on life expectancy and under-5 mortality rates for both sexes, stratified by gender and region. The dataset included numerical variables—life expectancy and under-5 mortality rates for both sexes, males, and females—across the years 2004 and 2024, allowing for a detailed examination of demographic trends.

We began by exploring the frequency distributions of these variables using histograms and central tendency measures, which showed distinct patterns across regions. For instance, Europe displayed the highest average life expectancy, while Africa showed the lowest. In contrast, Africa had the highest under-5 mortality rates, with Europe demonstrating the lowest, emphasizing stark regional disparities in health outcomes.

To assess the homogeneity of these variables, we performed an interquartile range (IQR) analysis on European sub-regions, which showed a homogeneous distribution of life expectancy in Northern and Southern Europe, while other sub-regions displayed more variability. Conversely, the under-5 mortality rates showed strong homogeneity across all examined sub-regions, suggesting uniform health conditions affecting child mortality in Europe. This analysis also highlighted notable heterogeneity between life expectancy figures across sub-regions, pointing to differences in regional life expectancy determinants.

Bivariate correlation analysis revealed significant associations, with a strong positive correlation among life expectancy measures across genders and a similarly strong positive correlation among under-5 mortality rates. A noteworthy finding was the strong negative correlation between life expectancy and under-5 mortality, indicating that improvements in one are linked to reductions in the other.

Finally, a comparison of variables between 2004 and 2024 highlighted a notable increase in life expectancy for both genders, alongside a substantial reduction in under-5 mortality rates. This indicates positive shifts in global health outcomes over the two decades, underscoring advancements in healthcare and living conditions.

While this analysis provided valuable insights, further research could involve hypothesis testing to explore causal factors underlying these trends. Additionally, a more granular examination of socio-economic or environmental factors could deepen the understanding of disparities in life expectancy and child mortality across regions.

Bibliography

P. Bhandari. correlation. https://www.investopedia.com/terms/c/correlation.asp/, Aug. 2022. Accessed: 27.10.2024.

P. Bhandari. Central tendency | understanding the mean, median mode. https://www.scribbr.com/statistics/central-tendency/, 2023. Accessed: 27.10.2024.

U. S. C. Bureau. Demographic research. https://www.census.gov/programs-surveys/international-programs/about/dem-soc-analysis.html, Jan. 2022. Accessed: 2023-04-11.

J. Frost. interquartile-range. https://statisticsbyjim.com/basics/interquartile-range//, May 2022. Accessed: 27.10.2024.

geeksforgeeks. Central tendency. https://www.geeksforgeeks.org/calculation-of-median-for-different-types-of-statistical-series/, Feb. 2022. Accessed: 27.10.2024.

D. M. Lane. *Introduction to Statistics*. Alpha Publications, 2001. Accessed: 31.10.2024.

D. Liberto. Central tendency. https://www.investopedia.com/terms/q/quartile.asp, Feb. 2022. Accessed: 27.10.2024.

S. Taylor. Central tendency. https://corporatefinanceinstitute.com/resources/data-science/central-tendency/, Feb. 2022. Accessed: 27.10.2024.

WallstreetMojo. Variance. https://www.wallstreetmojo.com/variance/#h-formula, 2021. Accessed: 27.10.2024.

S. Wilkes. *An Introduction to the Science of Statistics: From Theory to Implementation*. Addison-Wesley, 1994. Accessed: 31.10.2024.

M. Yi. Histogram in matplotlib. https://chartio.com/learn/charts/histogram-complete-guide/, June 2019. Accessed: 31.10.2024.